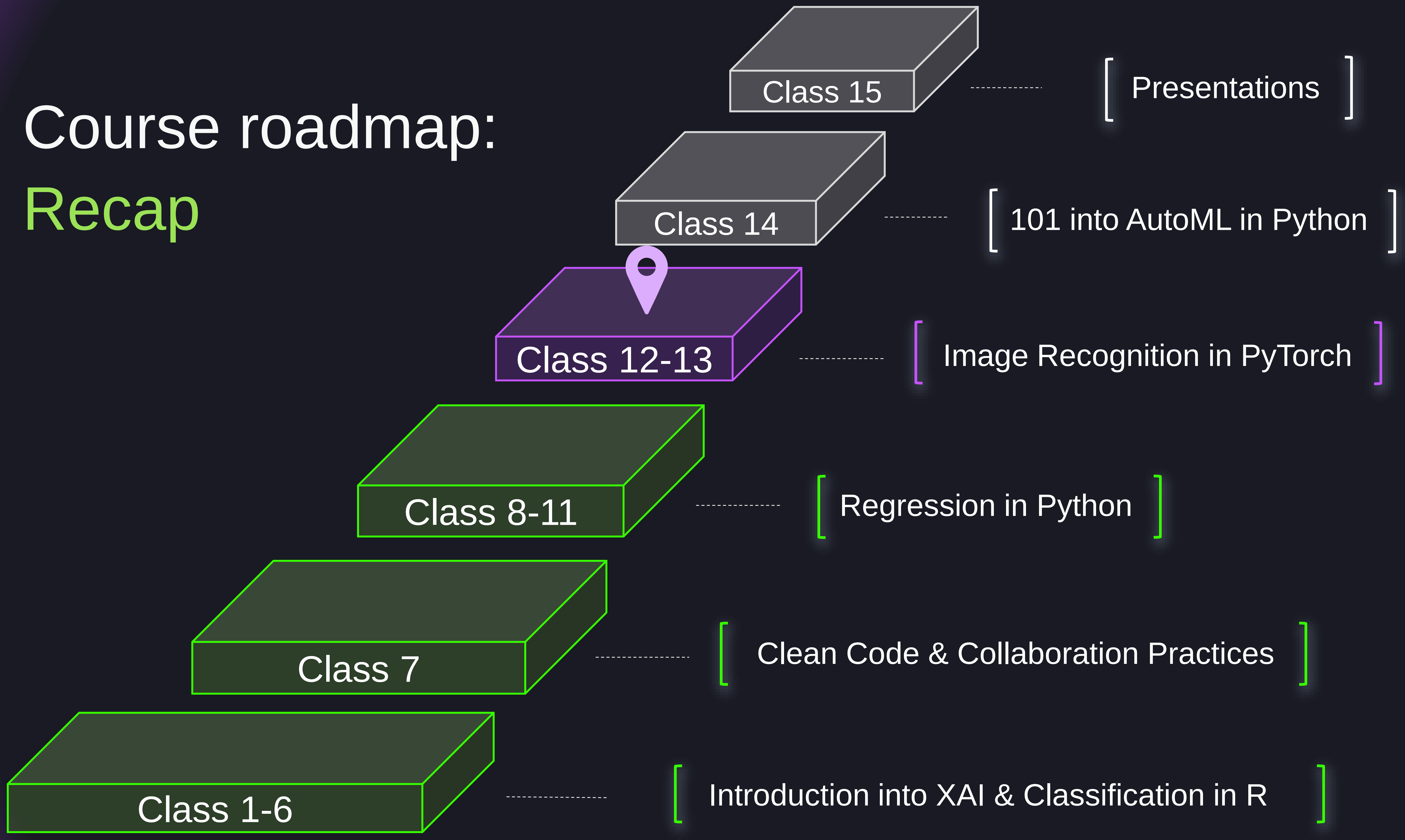
Faculty of Economic Sciences, UW

Applications of eXplainable AI in Predictive Modelling

Irena Zimovska

XAI 2025

Course roadmap:
Recap



Class 12-13

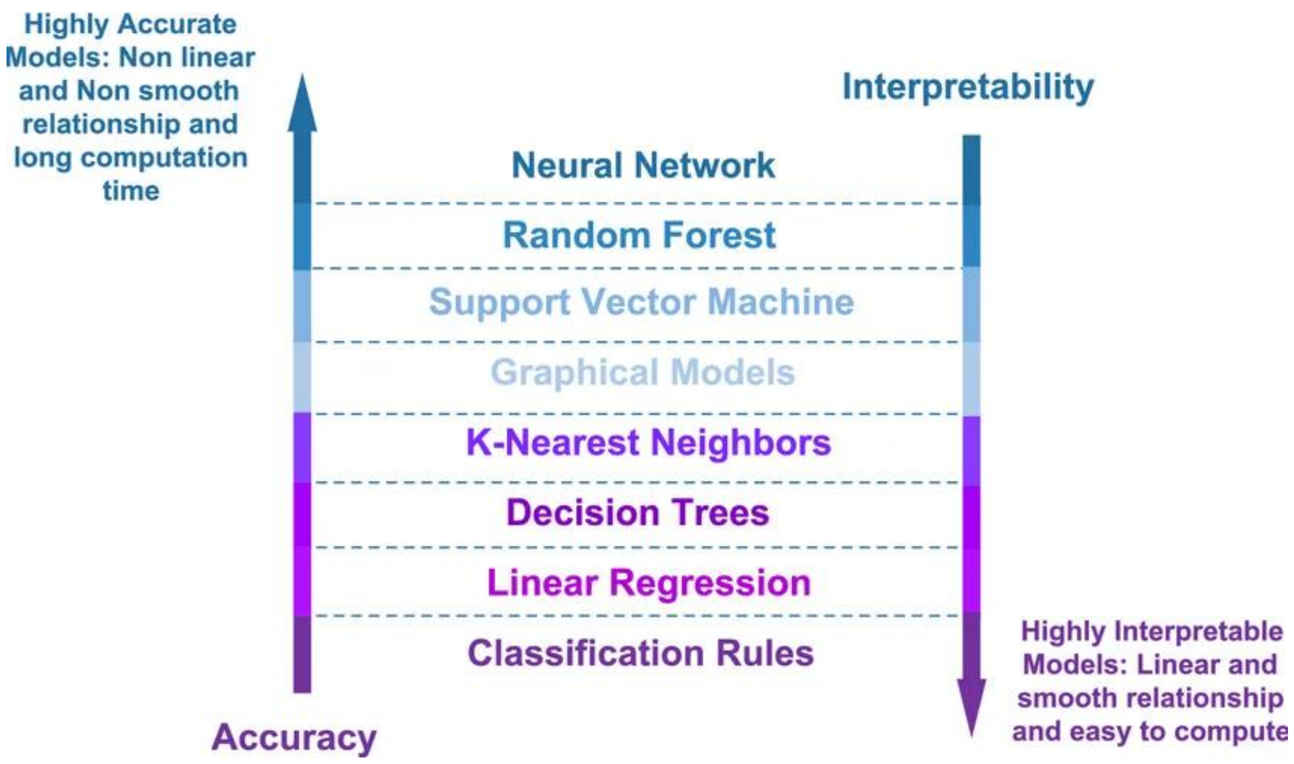
{ Image Recognition in PyTorch }

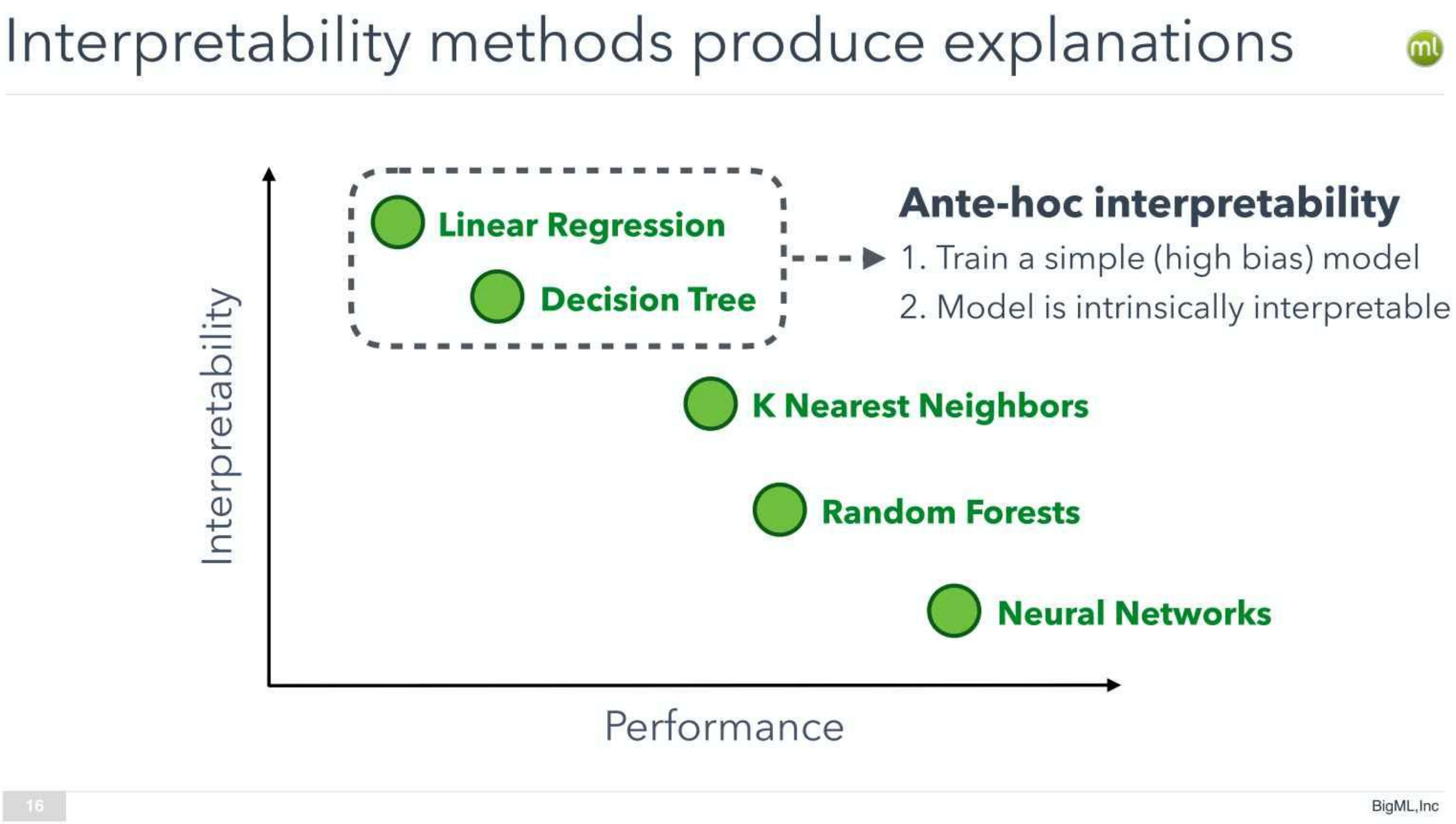
Irena Zimovska | WNE UW [XAI 2025]

- X Knowledge gained: XAI taxonomy
- X Tabular Data Explainers
- X Diving into Deep Neural Networks
- X Introduction to PyTorch
- X Image Data Explainers: SOTA

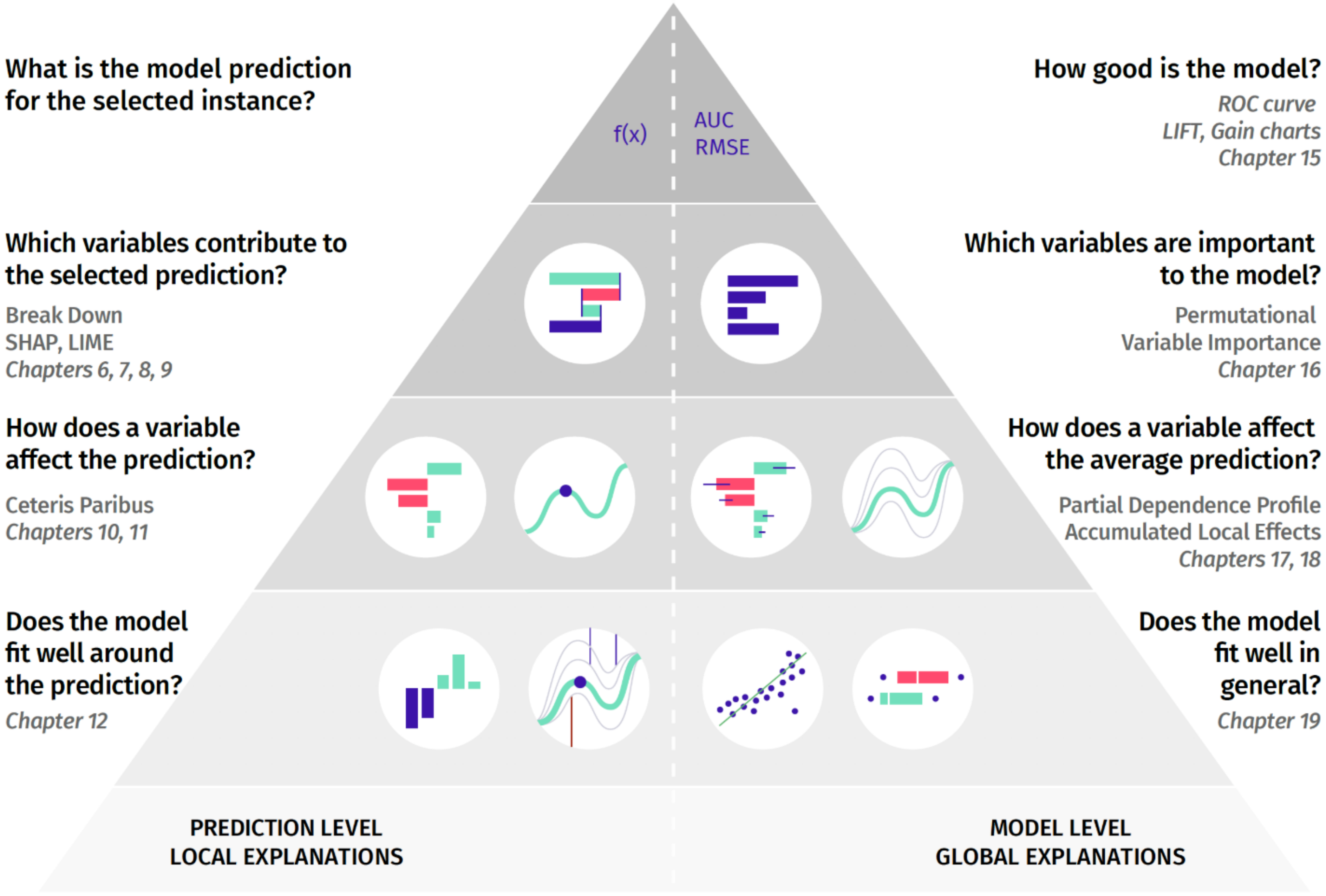
{ Knowledge Refinement }

Model Interpretability/Performance Trade-Off





Model Exploration Stack



Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis*. Chapman & Hall/CRC. https://pbiecek.github.io/ema/

XAI Brings Answers

Correctness: Are we sure that all, and only, the features of interest contributed to our algorithm's decisions?

Robustness: Are we sure the model is not susceptible to disturbances?

Bias: Are we aware of any specific biases in the data that unfairly penalise groups of individuals?

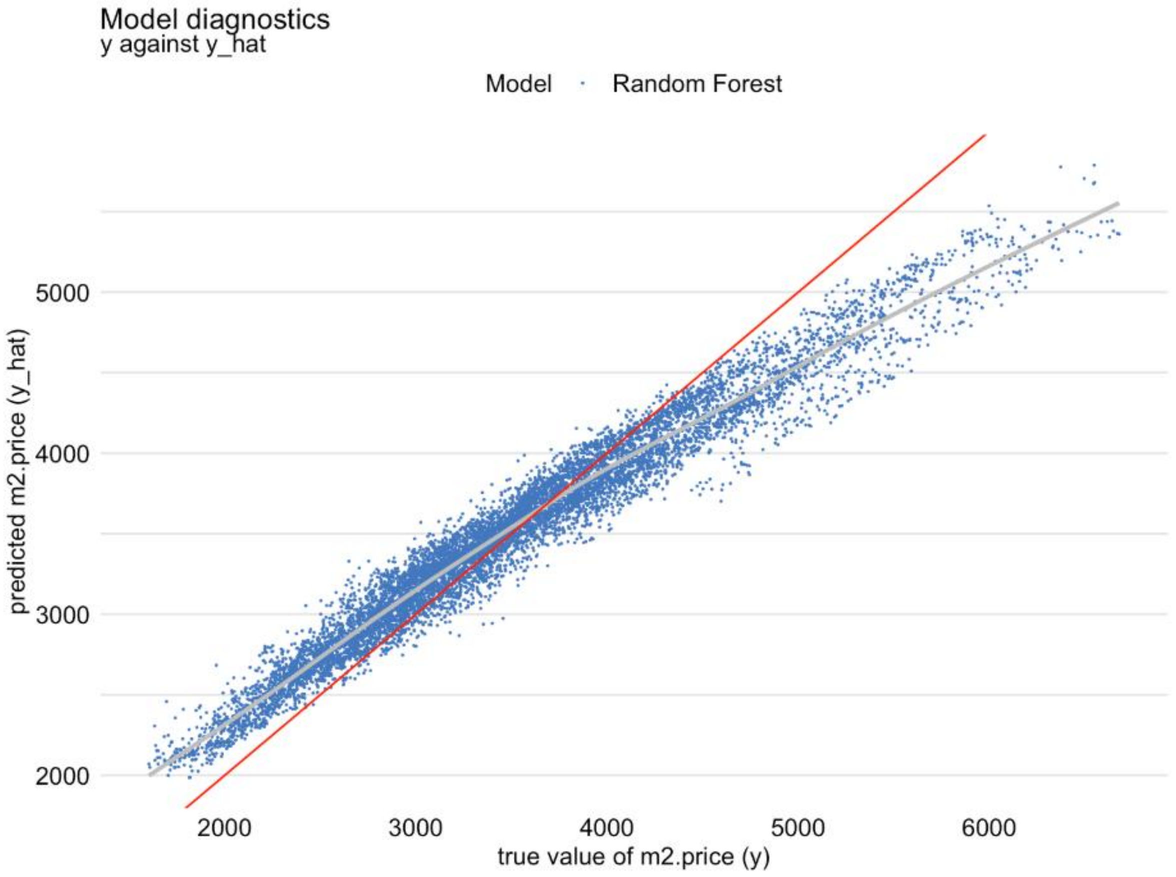
Improvement: In what specific way can the prediction model be improved?

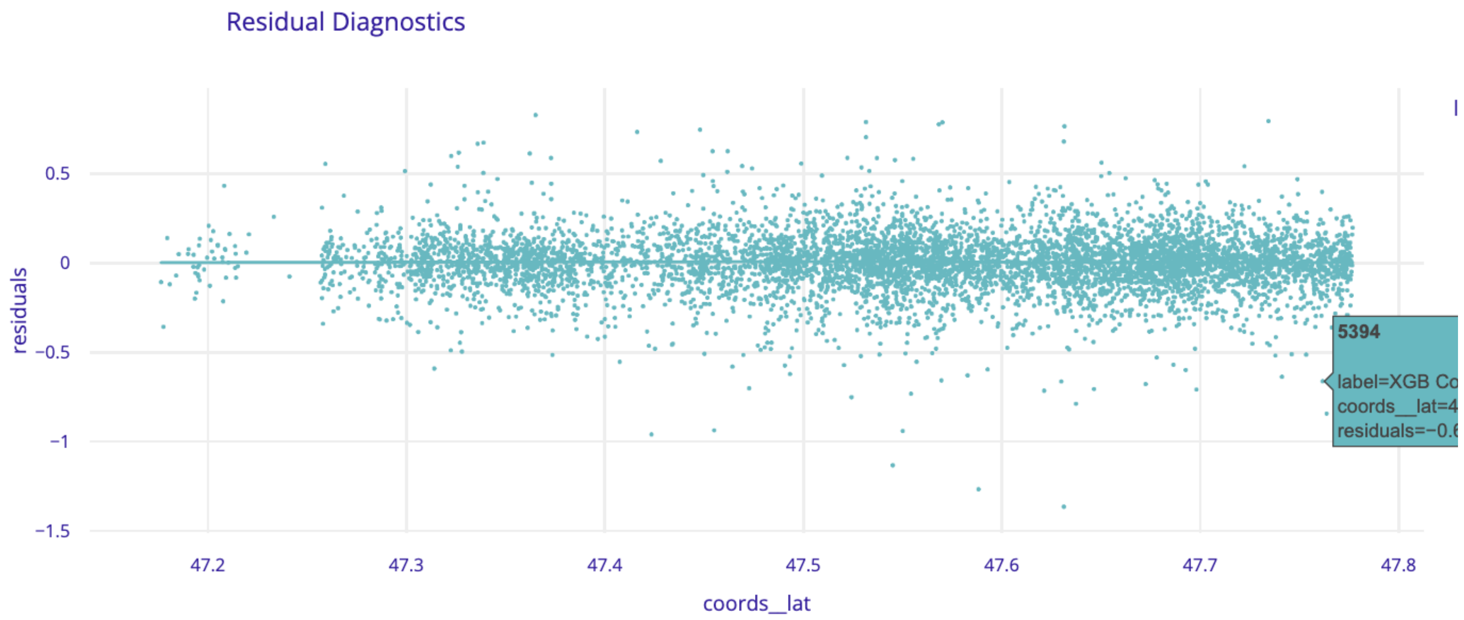
Transferability: Specifically how can the prediction model from one application domain be applied to another application domain?

Human understanding: Can we explain the model's algorithmic machinery to an expert or even to a layperson?

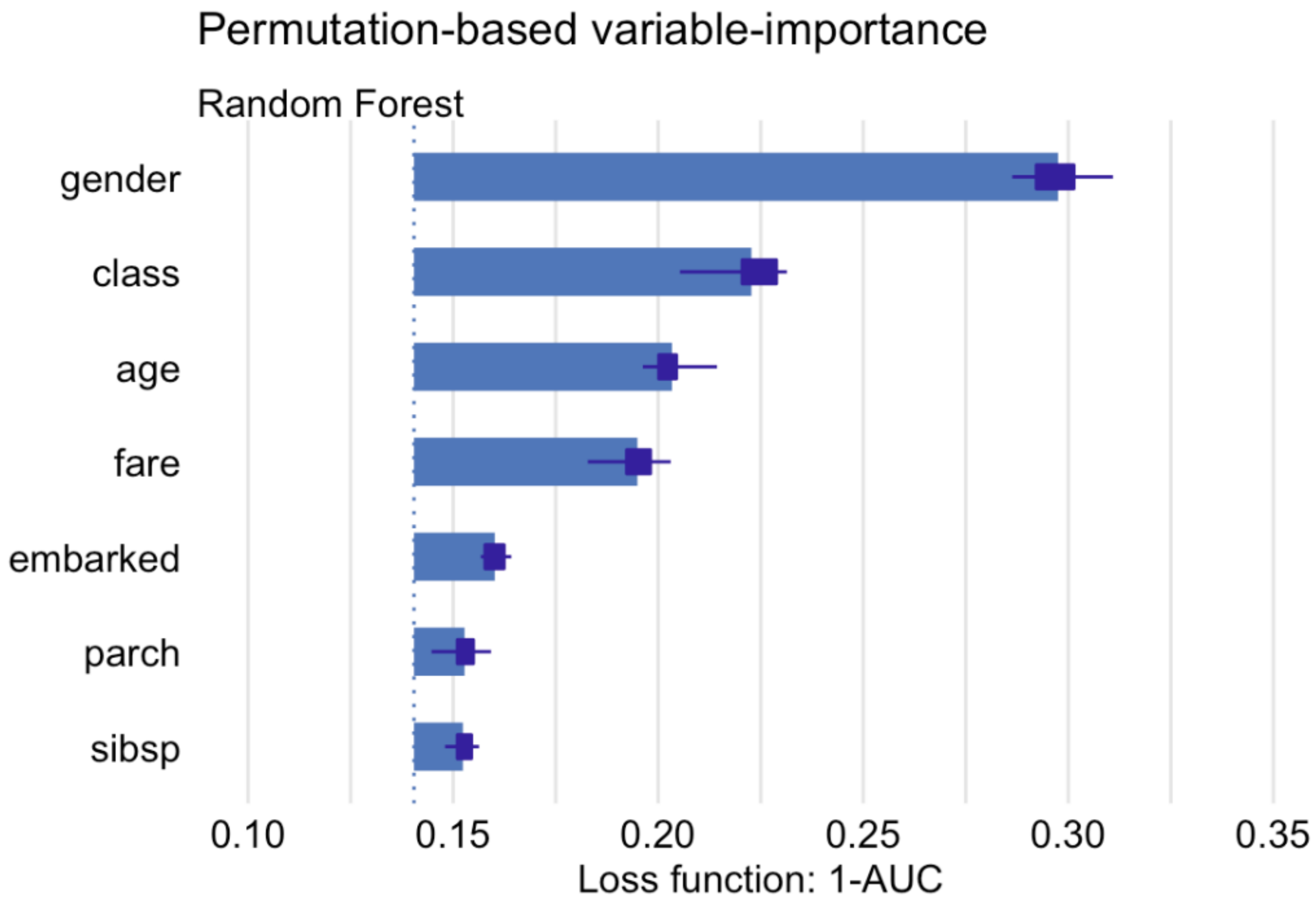
{ Explainers for Tabular Data }

Residual Diagnostics Plots





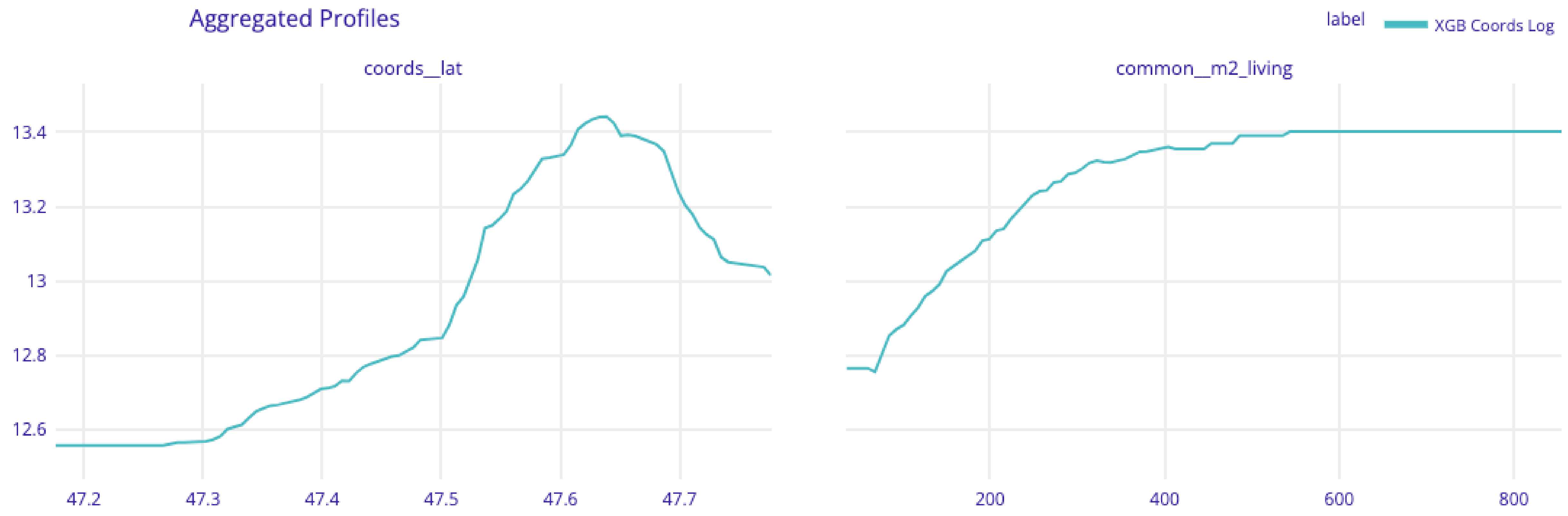
Permutation based feature importance



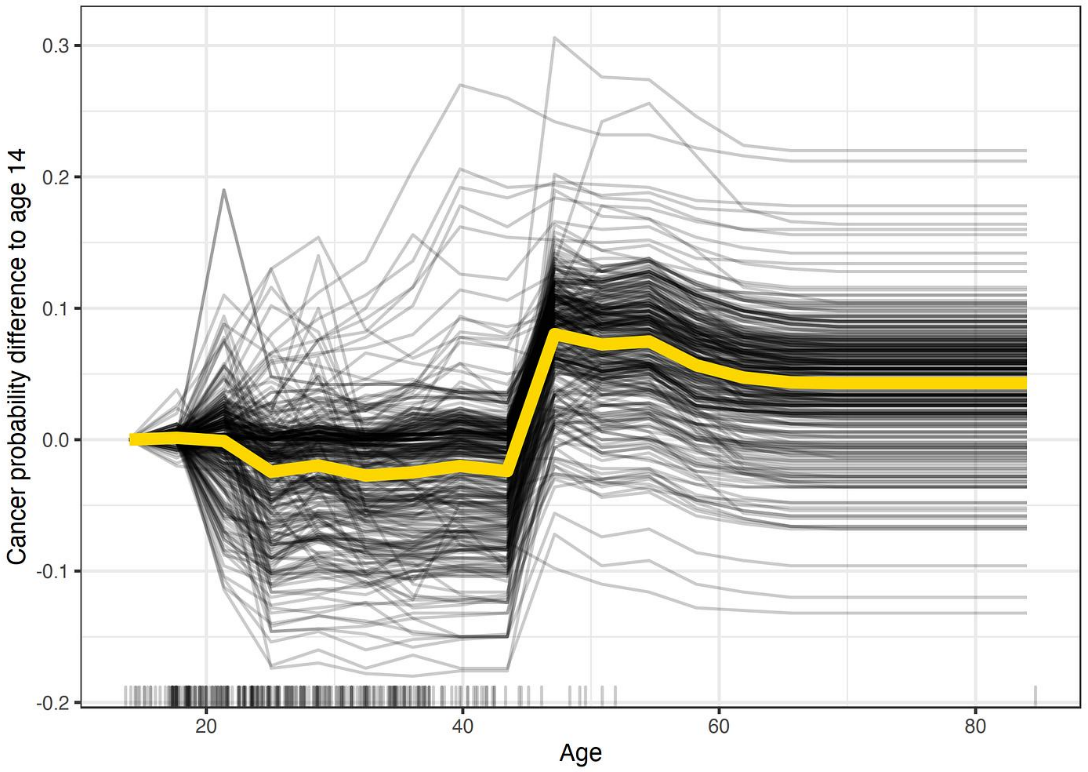
Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis*. Chapman & Hall/CRC. https://pbiecek.github.io/ema/

PDP (Partial Dependence Plot)

The PDP (Partial Dependence Plot) shows the marginal effect that one or two features have on the predicted result of a machine learning model.

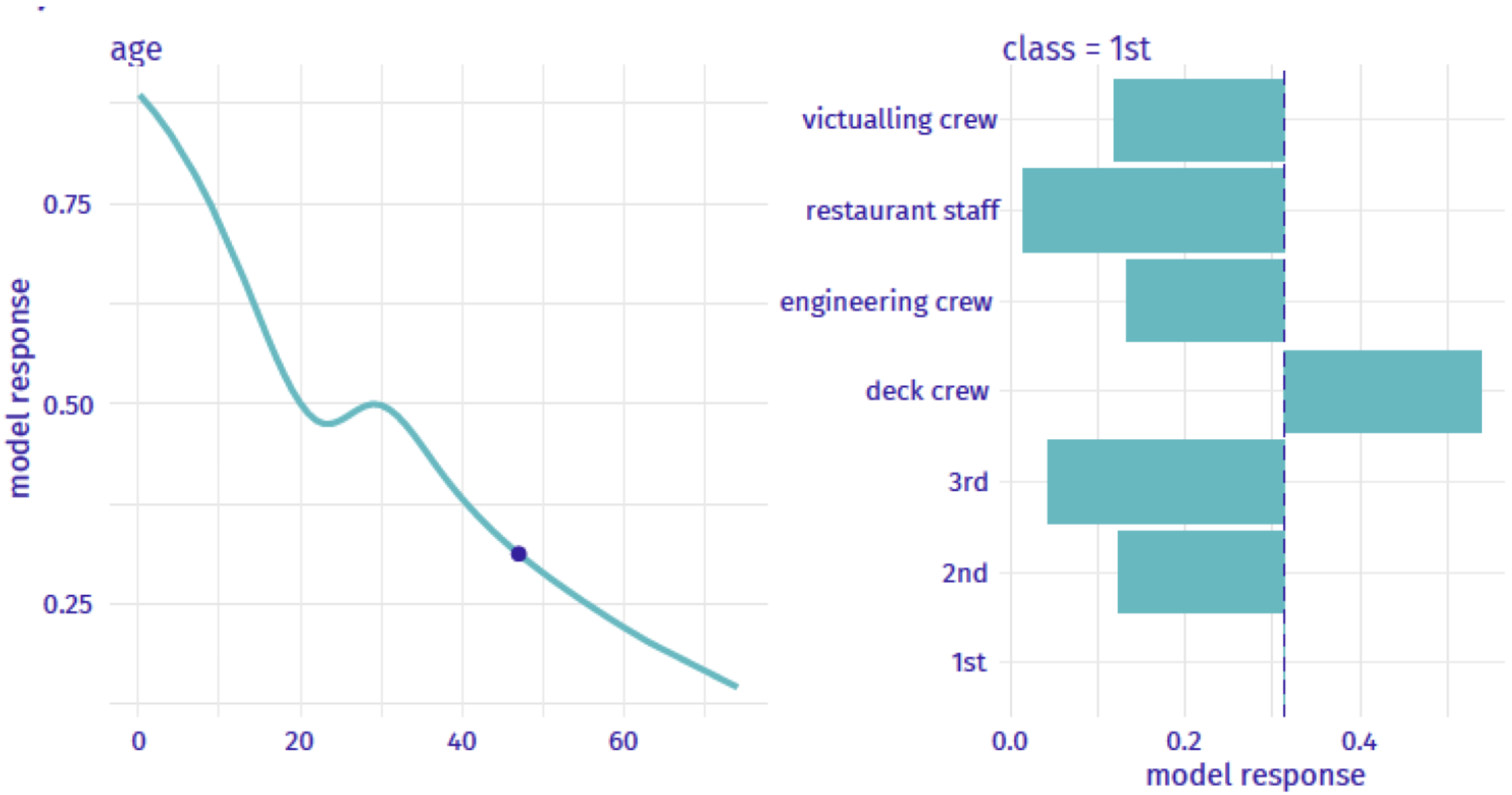


ICE (Individual Conditional Expectation)



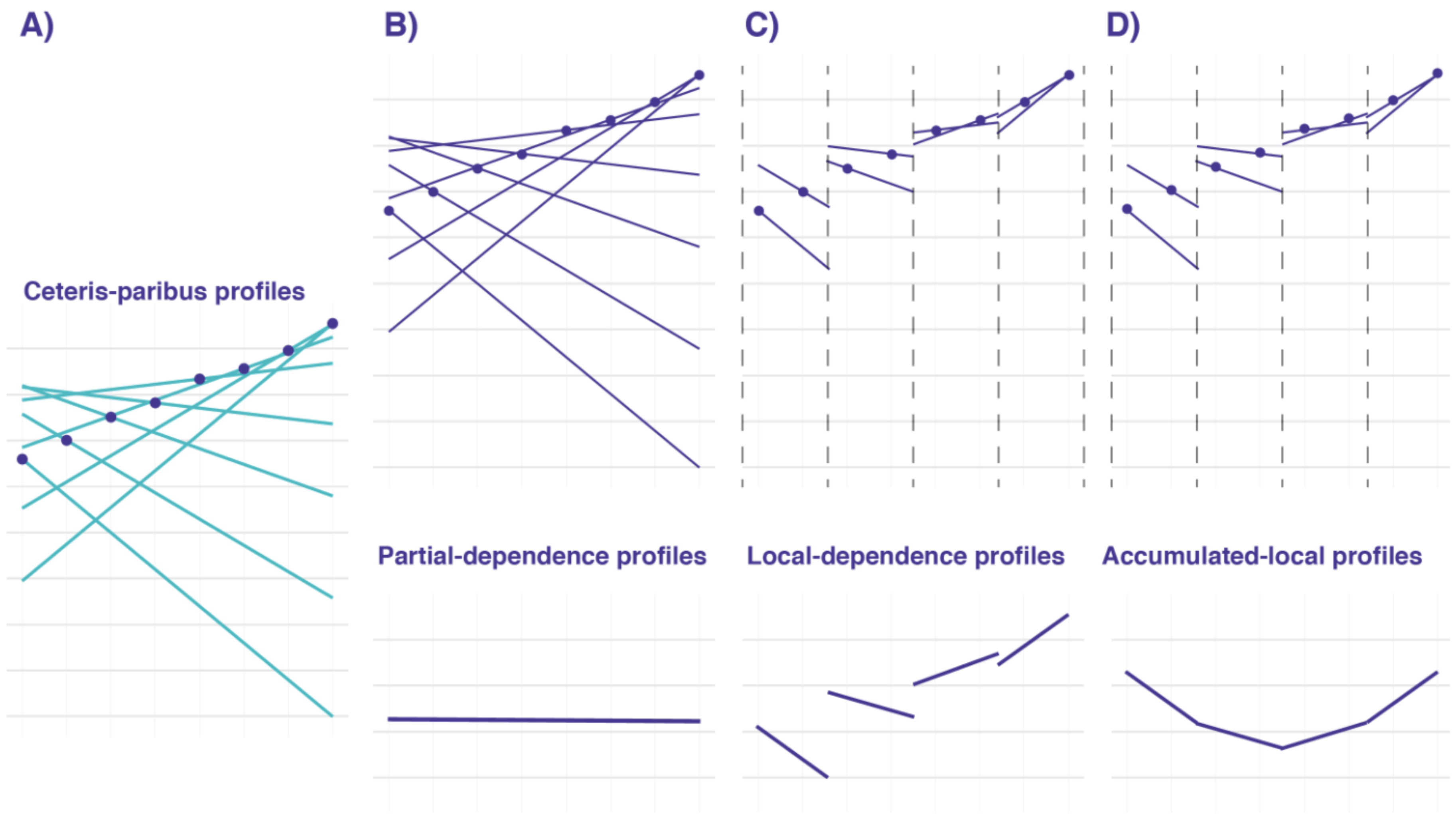
ICE (Individual Conditional Expectation) graphs show how the prediction of the instance changes when a feature changes. The Partial Dependence plot for the average effect of a feature is a global method because it does not focus on specific cases, but instead on a global average. The equivalent of a PDP for individual data instances is called an Individual Conditional Expectation (ICE) plot.

CPP (Ceteris Paribus Profiles)



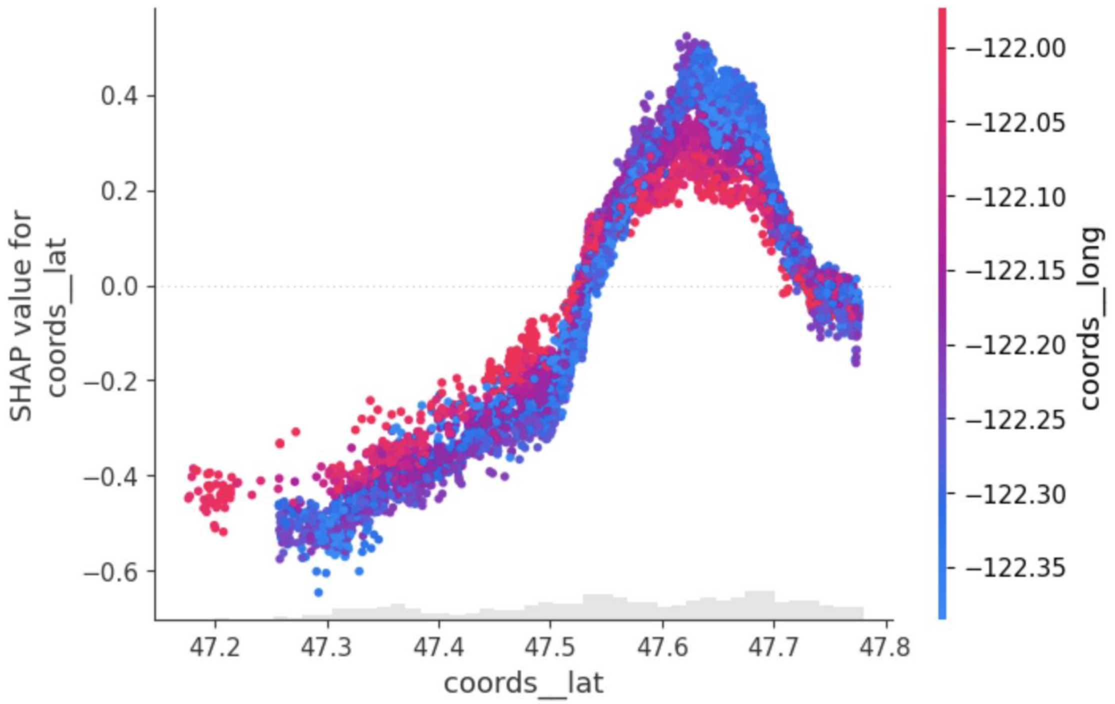
Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis*. Chapman & Hall/CRC. https://pbiecek.github.io/ema/

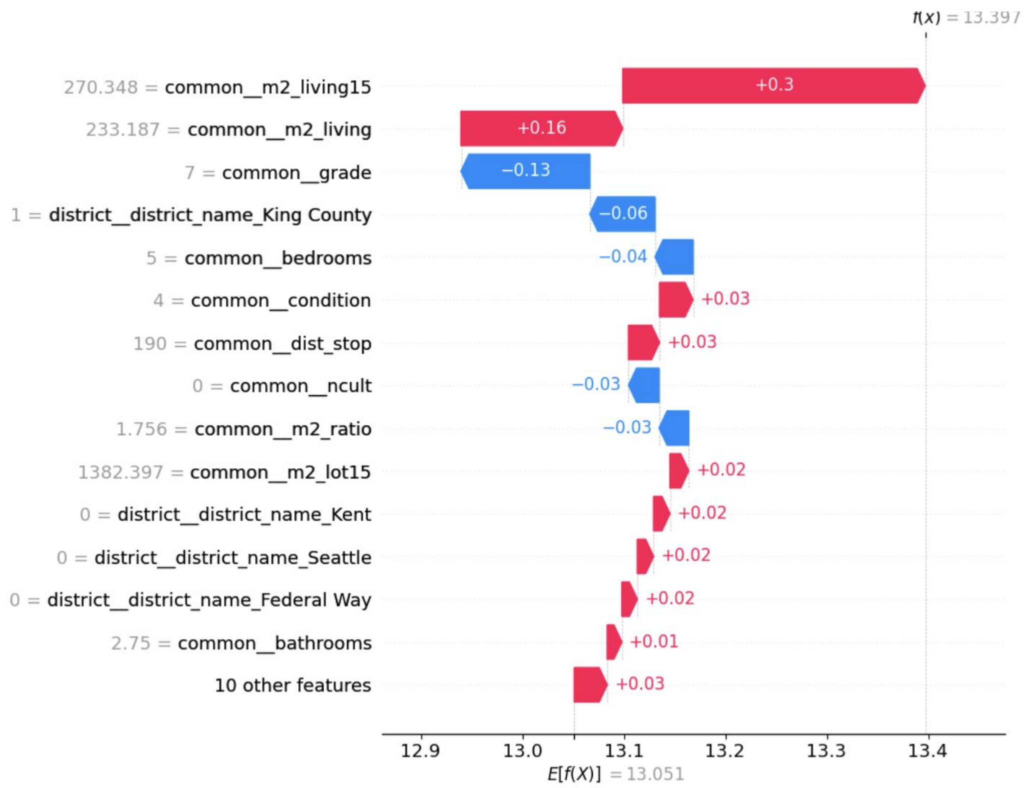
ALE (Additive Local Effects)



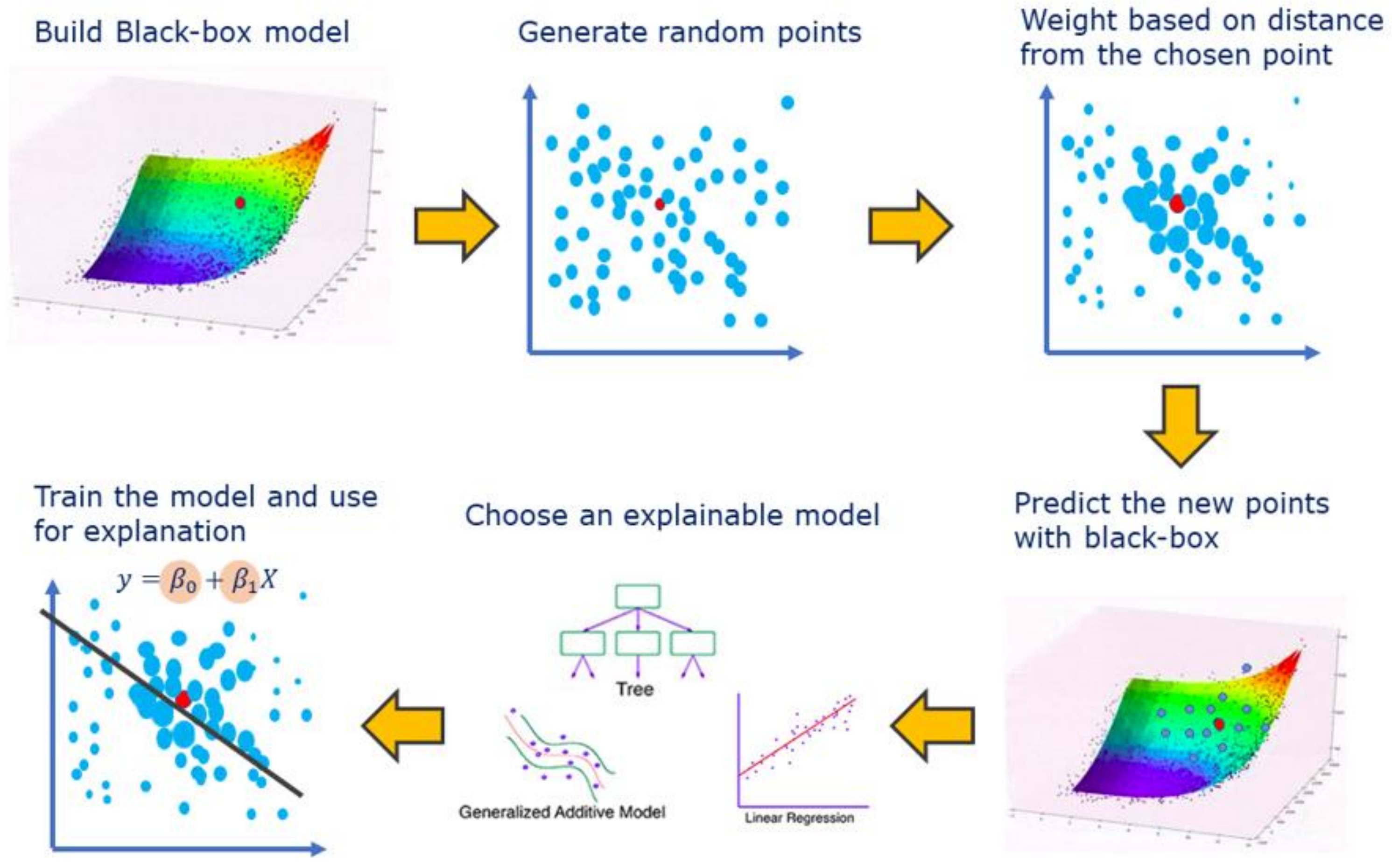
Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis*. Chapman & Hall/CRC. https://pbiecek.github.io/ema/

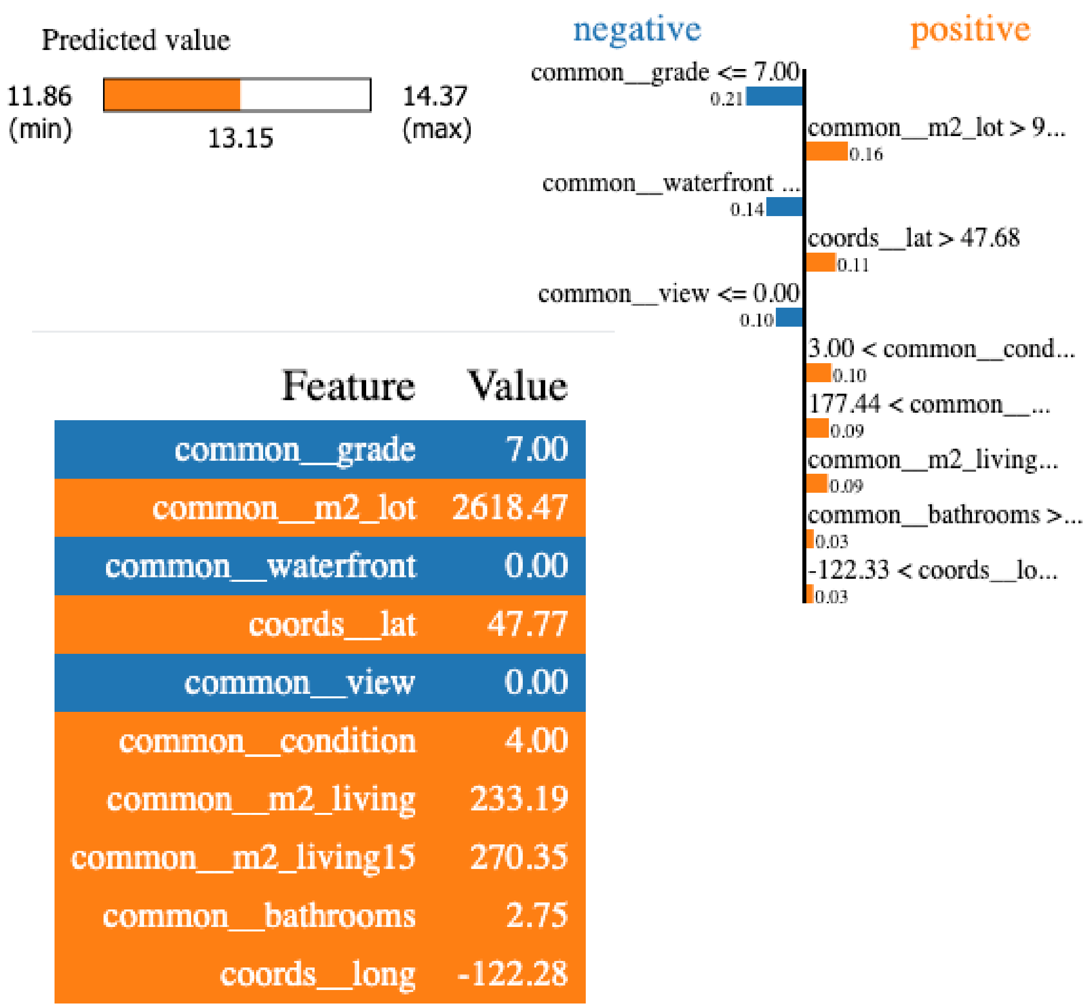
SHAP (Shapley Additive exPlanations)





LIME

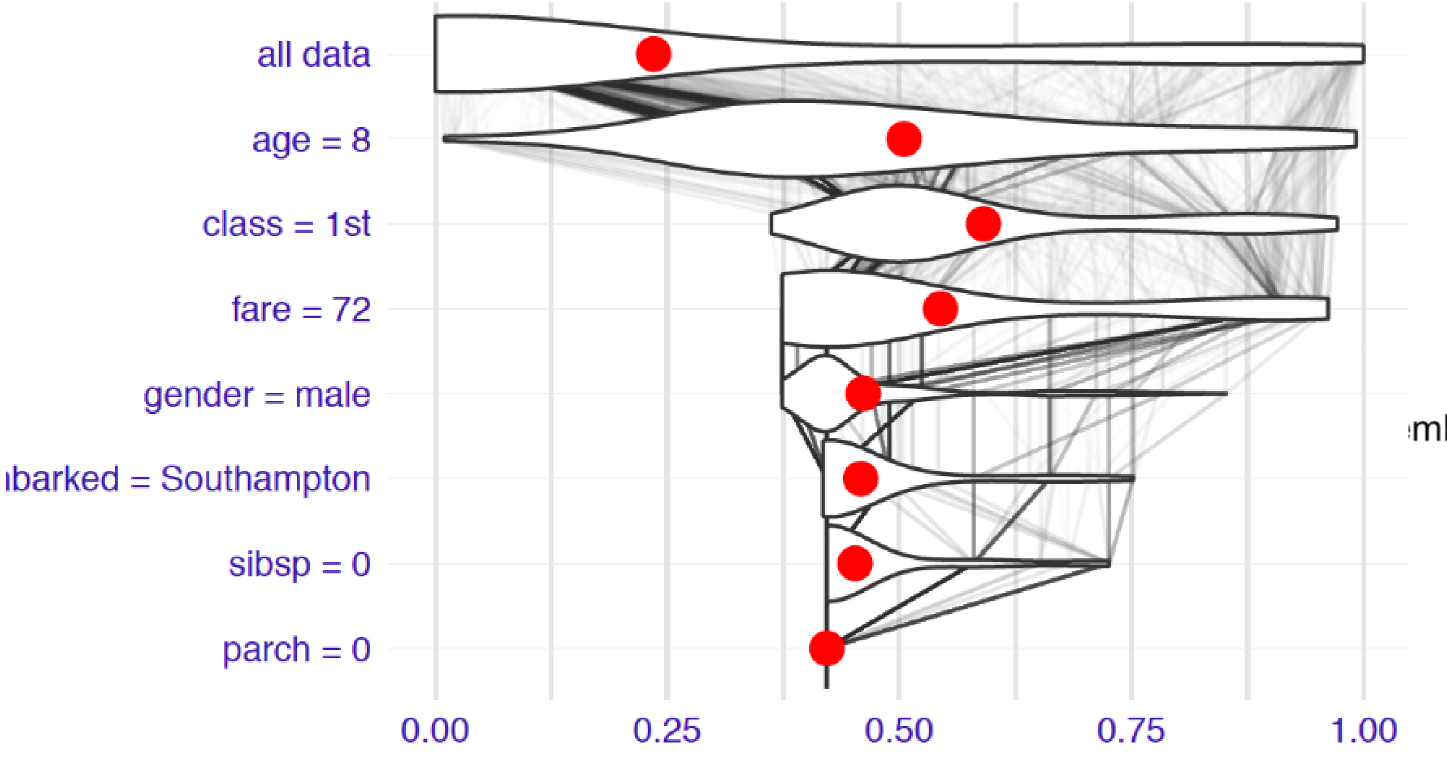


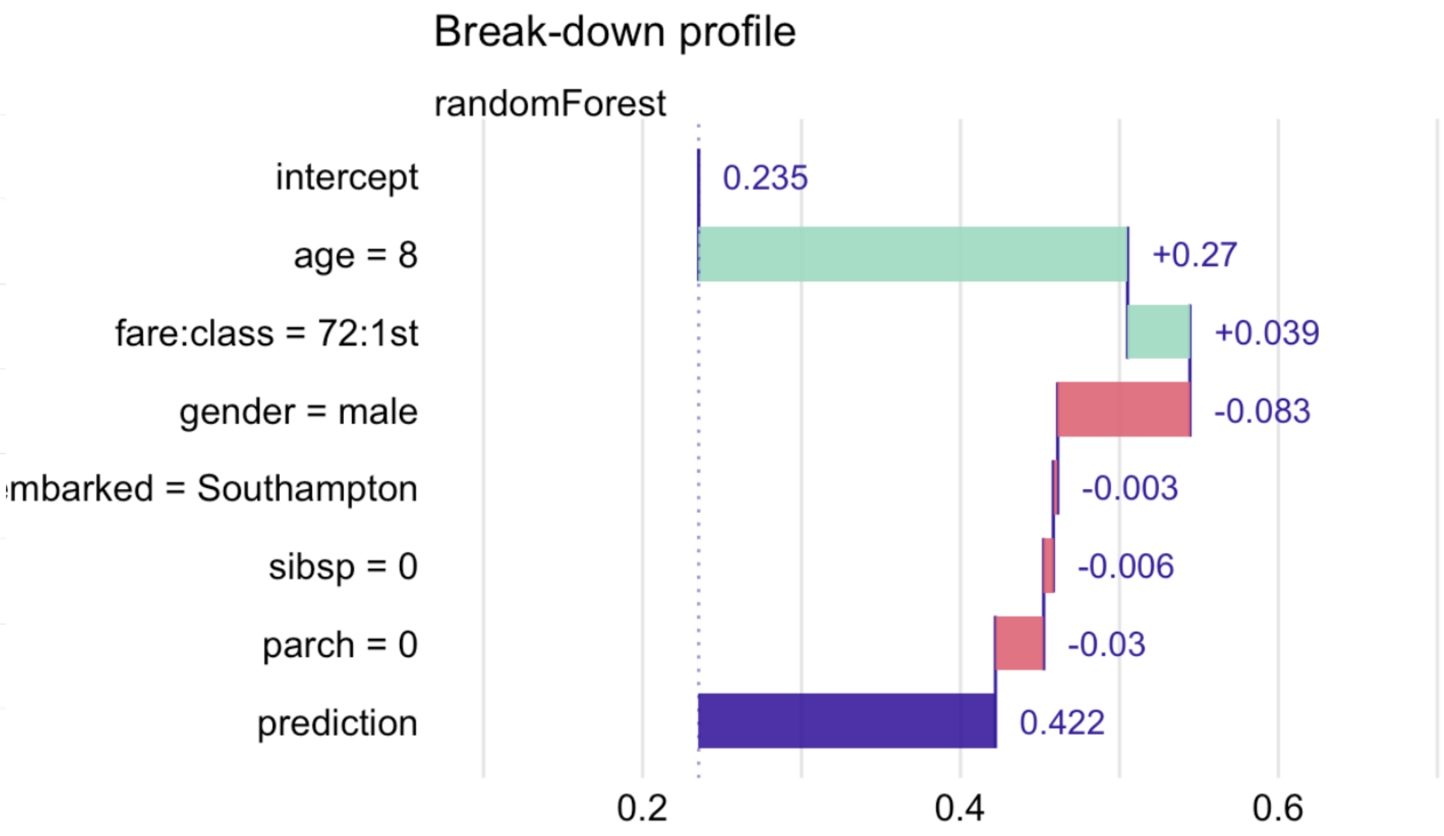


Feature	Value
common__grade	7.00
common__m2_lot	2618.47
common__waterfront	0.00
coords__lat	47.77
common__view	0.00
common__condition	4.00
common__m2_living	233.19
common__m2_living15	270.35
common__bathrooms	2.75
coords__long	-122.28

https://medium.com/@shreeraj260405/lime-unveiled-a-deep-dive-into-explaining-ai-models-for-text-images-and-tabular-data-046c7c3b4e9f

Break Down Plots





Frameworks Available in Python

DALEX

SHAP

ALIBI EXPLAIN

InterpretML

LIME

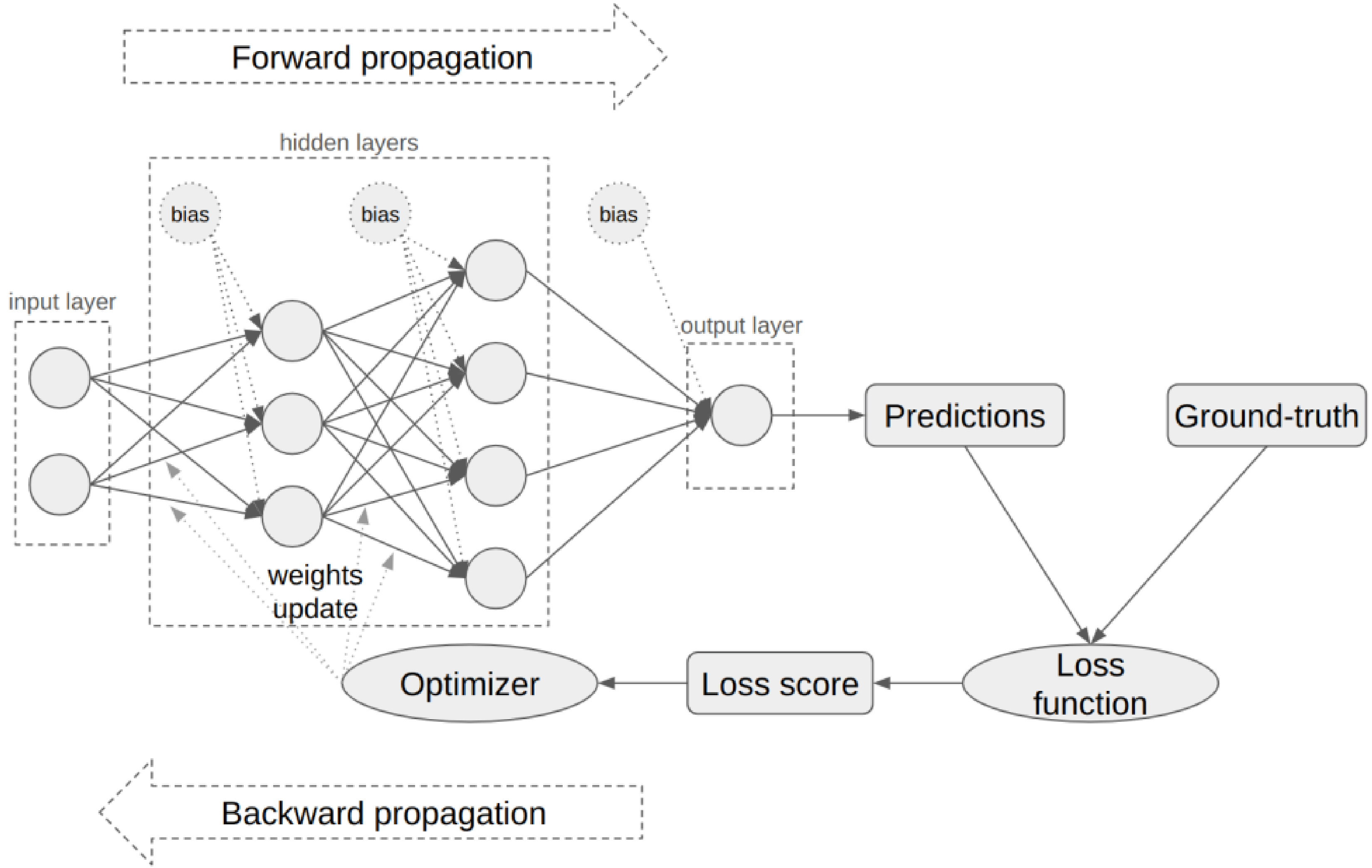
Captum

OmniXAI

iNNvestigate

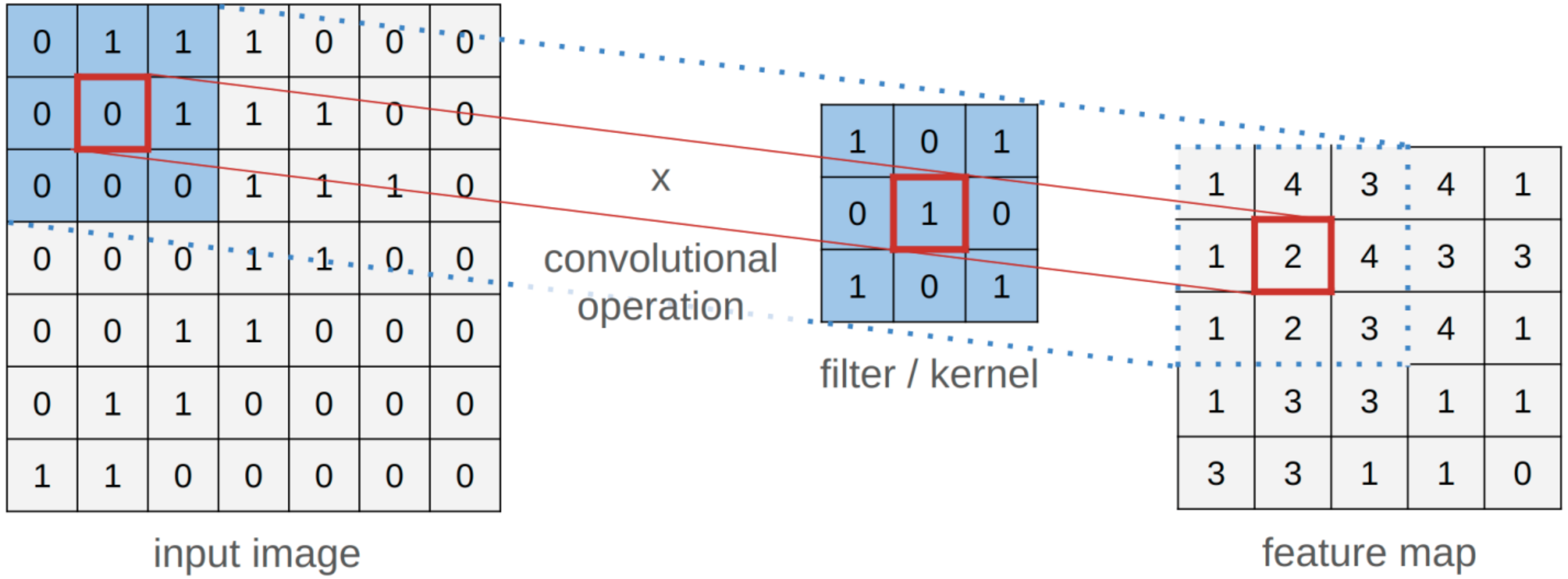
03 { Deep Neural Networks Revisited }

Network Architecture



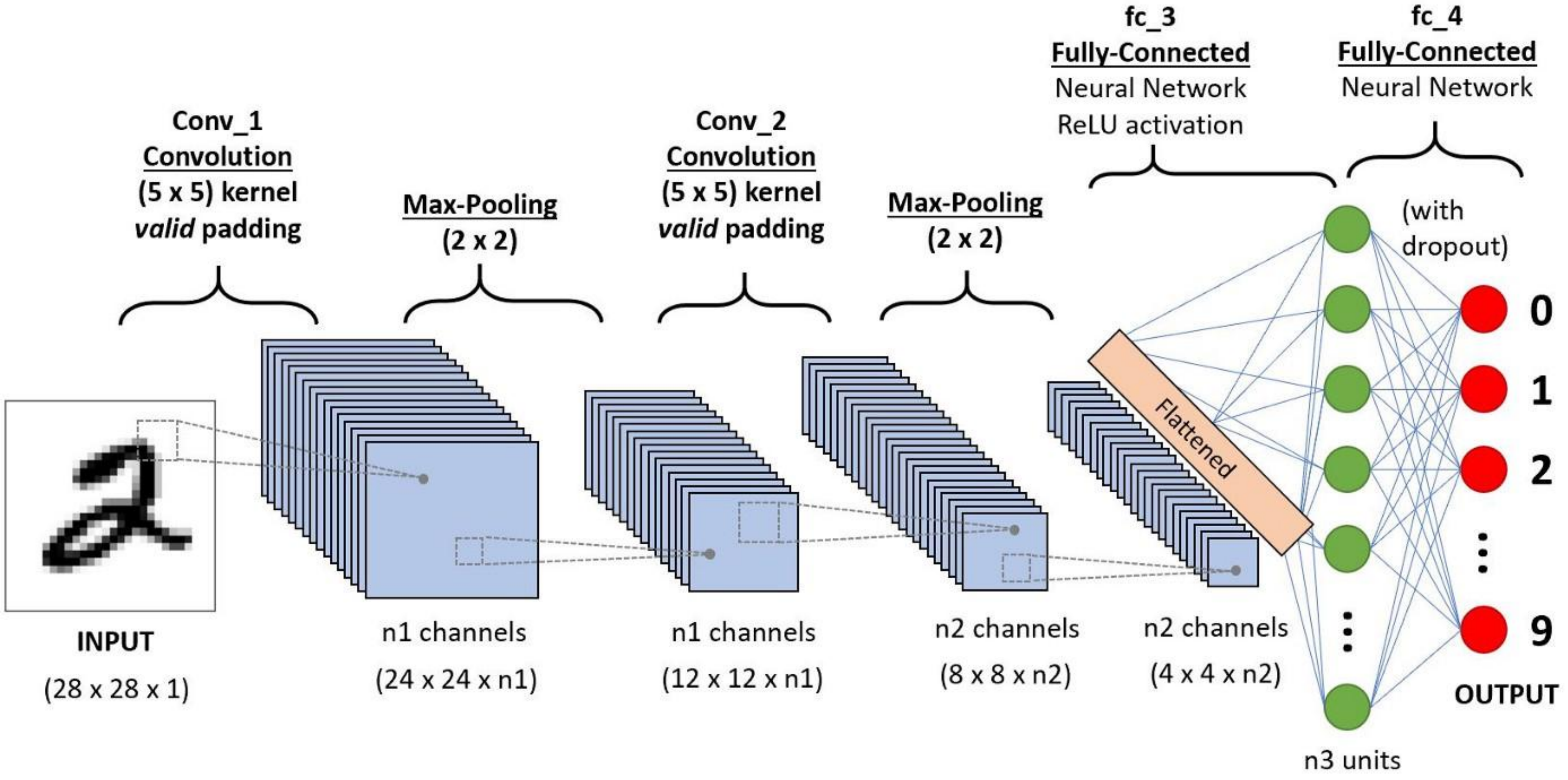
Hryniewska-Guzik, W. (2024). A multi-level perspective on the deep learning models and human-oriented explanations with applications to medical images (Doctoral dissertation). Warsaw University of Technology, Warsaw, Poland.

Convolution Operation



Hryniewska-Guzik, W. (2024). A multi-level perspective on the deep learning models and human-oriented explanations with applications to medical images (Doctoral dissertation). Warsaw University of Technology, Warsaw, Poland.

Convolutional Neural Network Architecture



Check out layers in PyTorch: https://docs.pytorch.org/docs/stable/nn.html#vision-layers

04 { Intro to PyTorch }

Installation

Start Locally

Select your preferences and run the install command. Stable represents the most currently tested and supported version of PyTorch. This should be suitable for many users. Preview is available if you want the latest, not fully tested and supported, builds that are generated nightly. Please ensure that you have **met the prerequisites below (e.g., numpy)**, depending on your package manager. You can also install previous versions of PyTorch. Note that LibTorch is only available for C++.

NOTE: Latest PyTorch requires Python 3.9 or later.

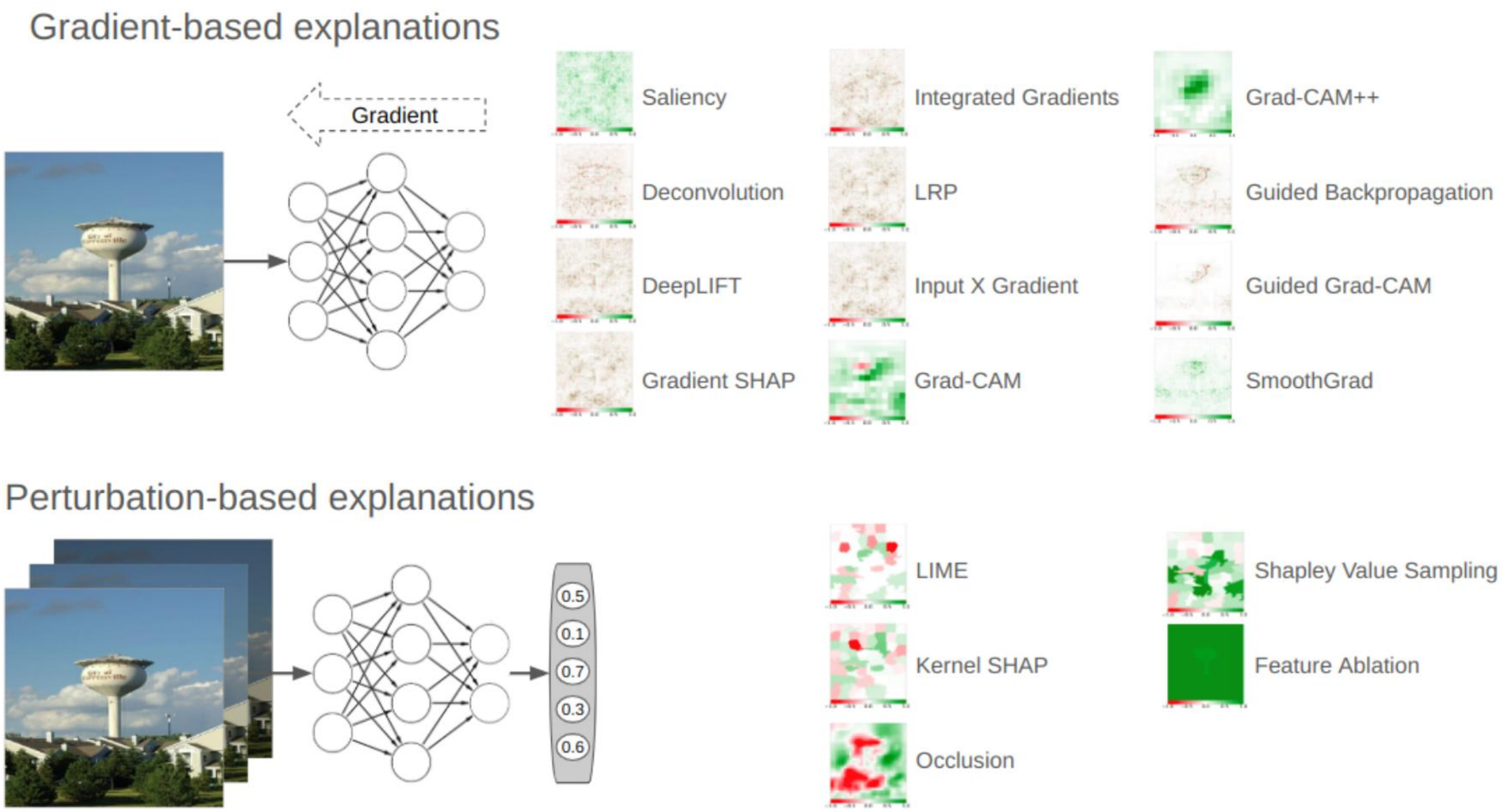
PyTorch Build	Stable (2.7.0)			Preview (Nightly)	
Your OS	Linux	Mac		Windows	
Package	Pip	LibTorch		Source	
Language	Python			C++ / Java	
Compute Platform	~~CUDA 11.8~~	~~CUDA 12.6~~	~~CUDA 12.8~~	~~ROCm 6.3~~	Default
Run this Command:	`pip3 install torch torchvision torchaudio`				

https://pytorch.org/get-started/locally/

Time to get back to code :)

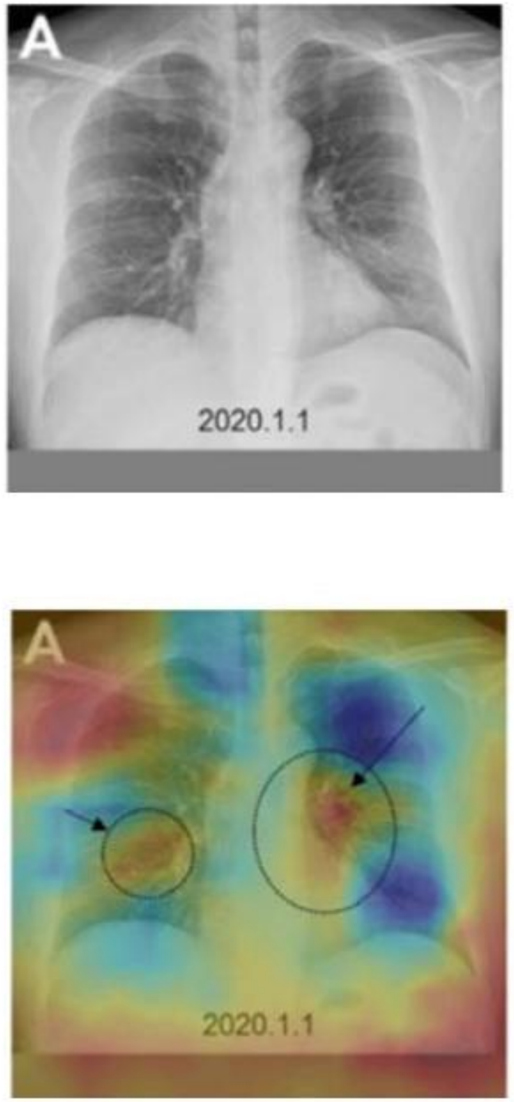
{ Explainers for Image Data }

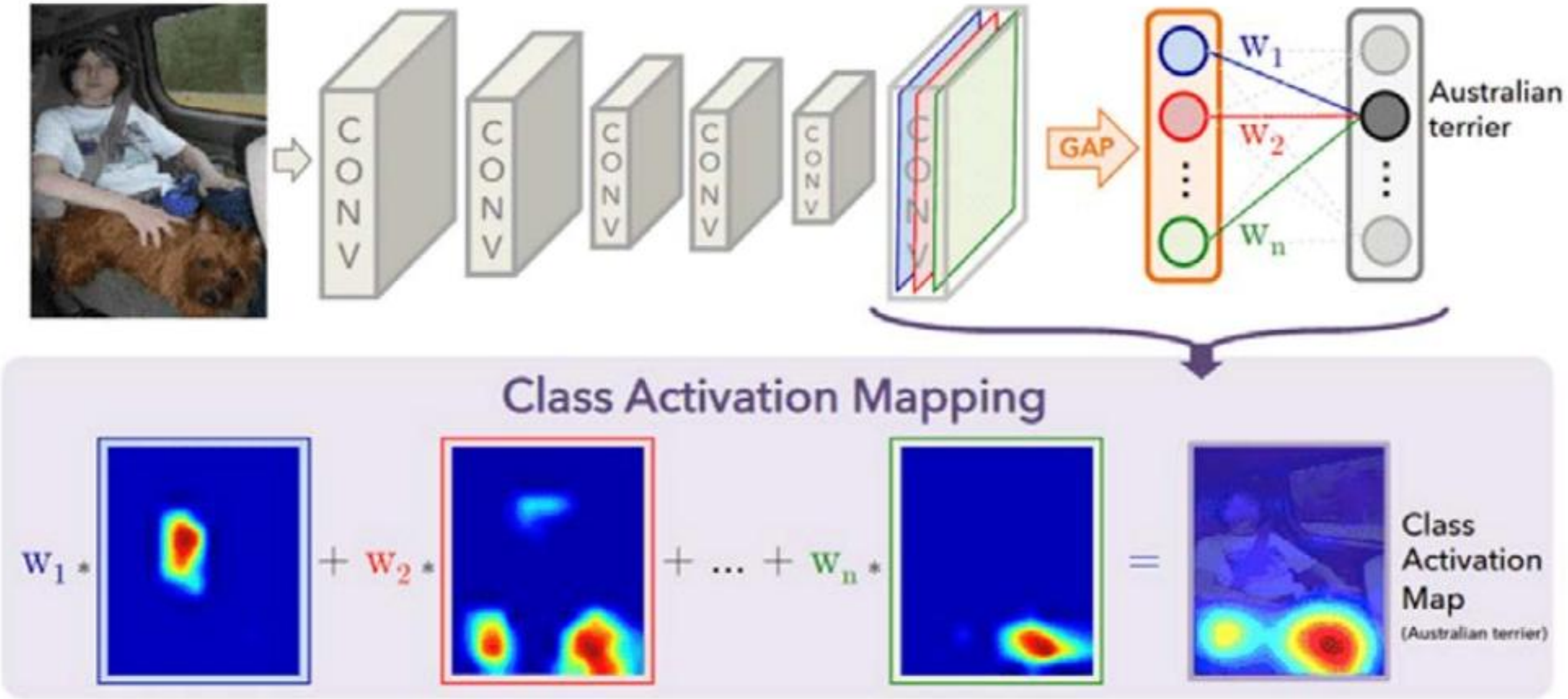
Local, Post-Hoc explainers



Hryniewska-Guzik, W. (2024). A multi-level perspective on the deep learning models and human-oriented explanations with applications to medical images (Doctoral dissertation). Warsaw University of Technology, Warsaw, Poland.

CAM (Class Activation Mapping)





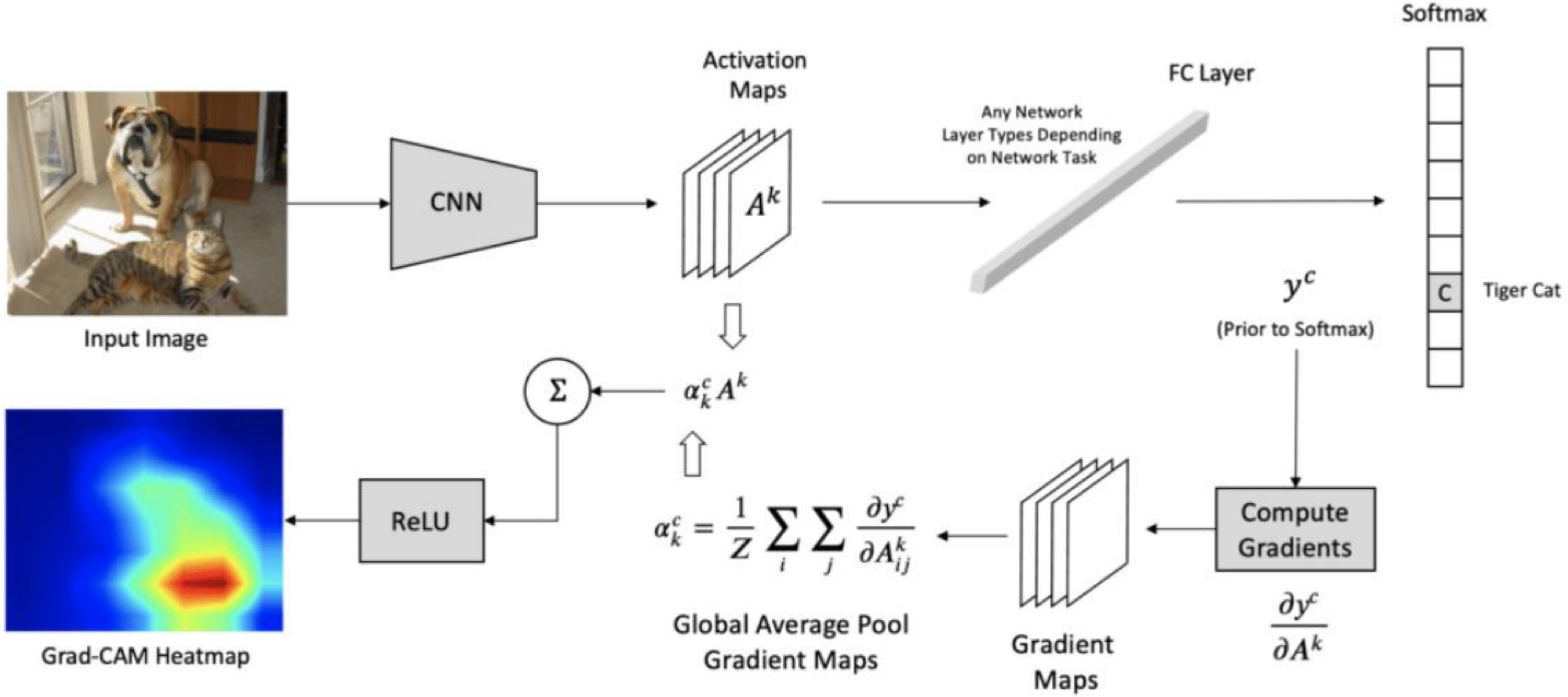
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localisation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2921-2929, https:// doi.org/10.1109/CVPR.2016.319

Grad-CAM (Gradient Class Activation Mapping)

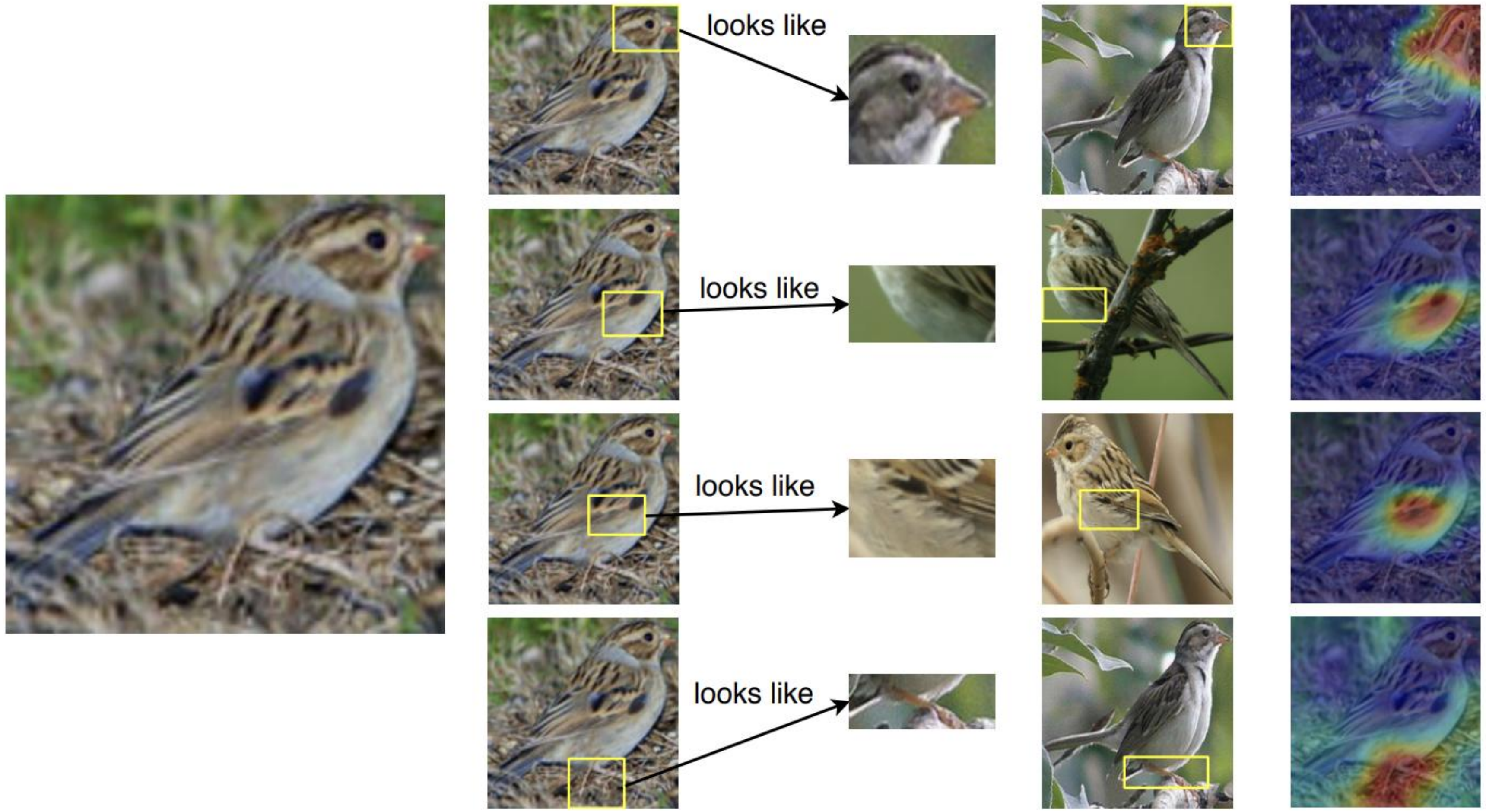
Grad-CAM (Gradient-weighted Class Activation Mapping) is an extension based on CAM, which uses the gradients for the target class that derives in the final convolutional layer. Unlike CAM, this method does not require any retraining and is broadly applicable to any architecture based on convolutional neural networks (CNN).

First, the class score gradient is calculated for the activation maps in the last convolutional layer. The gradients are returned after averaging them over the size of the activation map, and then the importance weights are calculated.

Grad-CAM (Gradient Class Activation Mapping)



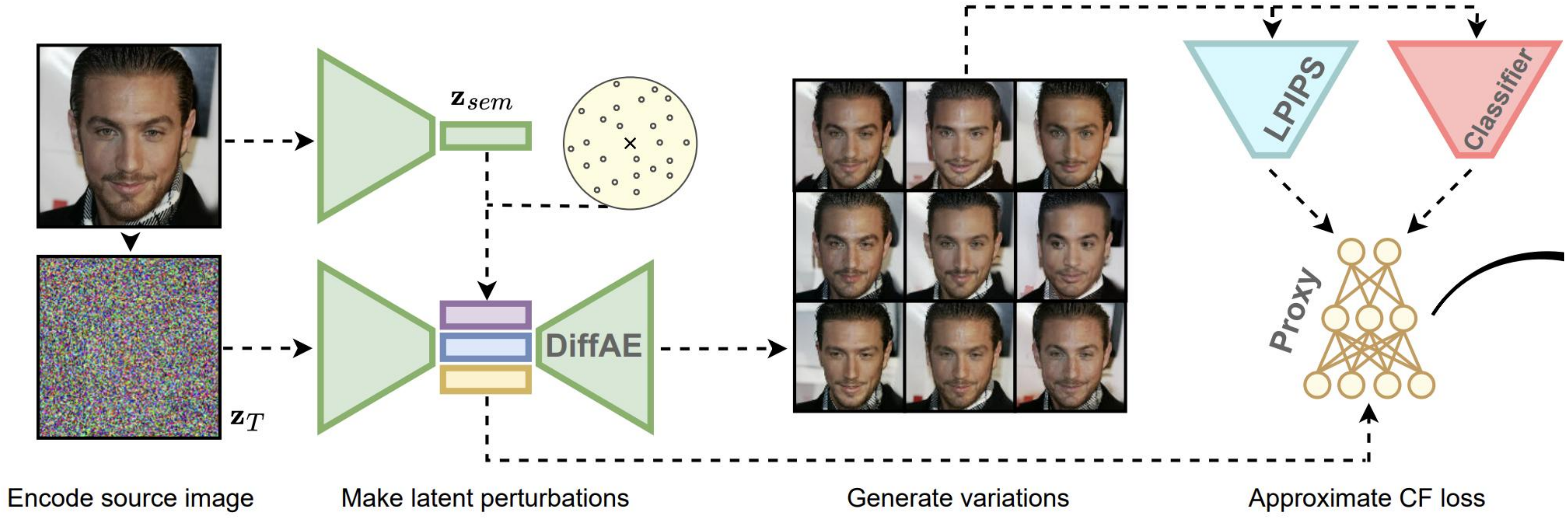
Prototypical Explanations



Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look simila to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical parts learned by our algorithm
Fourth column: source images of the prototypical parts in the third column
Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems, 32.

Counterfactual Explanations



Sobieski, Bartlomiej, and Przemyslaw Biecek. "Global counterfactual directions." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.